

Real-time detection of transients in OGLE-IV with application of machine learning

Jakub Klencki¹ and Łukasz Wyrzykowski¹

1. Warsaw University Astronomical Observatory
Al. Ujazdowskie 4, 00-478 Warsaw, Poland

The current bottleneck of transient detection in most surveys is the problem of rejecting numerous artifacts from detected candidates. We present a triple-stage hierarchical machine learning system for automated artifact filtering in difference imaging, based on self-organizing maps. The classifier, when tested on the OGLE-IV Transient Detection System, accepts 97% of real transients while removing up to 97.5% of artifacts.

1 Introduction

The task of analyzing the observational data is especially challenging for time-domain surveys searching for transients (such as OGLE (Udalski et al., 2015)), as they usually attempt to perform data analysis as close to real-time as possible. Detecting an ongoing transient as early as possible not only allows for triggering extra follow-up observations, leading to far better light curve coverage, but also gives a chance to observe the earliest stages of the event - possibly essential for understanding the physics of the phenomena. Machine learning techniques are a promising approach to help solve this problem by performing data classification automatically (Bloom & Richards, 2012). In our work we aim to further develop the OGLE-IV Transient Detection System (Wyrzykowski et al., 2014) by automatically rejecting numerous bogus detections - the most time consuming stage of transient detection.

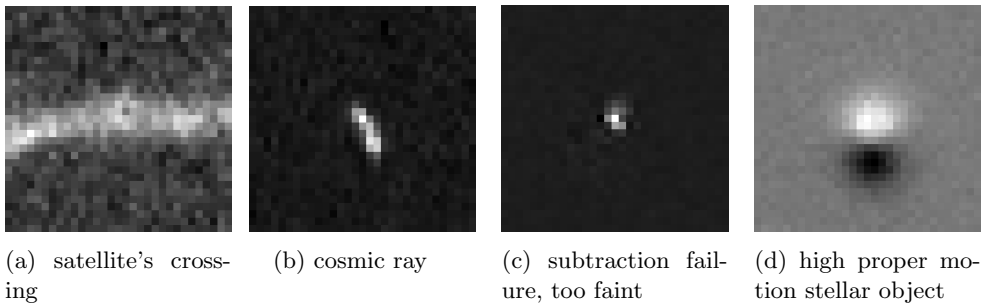


Fig. 1: Most common types of artifacts from the OGLE transient pipeline. Presented as 31x31 pixel cutouts from image differences.

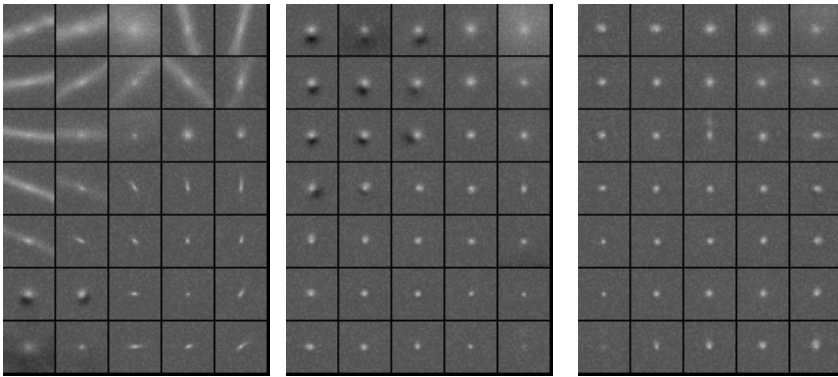
The OGLE pipeline identifies candidates for transients using Difference Imaging Analysis and cuts the subtraction images into thumbnails centered around positive subtraction residuals. The pipeline produces thousands of candidates each night, the vast majority of which being unwanted artifacts and false subtractions – we call them bogus detections (see Fig.1).

2 Method: Self-Organizing Maps

Self-Organizing Maps (SOM) (Kohonen, 1982) is an artificial neural network, usually in a form of 2D array of nodes with weight vectors associated to them. Each analyzed datum is represented in a form of a numeric vector and mapped onto node with weight vector most similar to it – the winner node.

The SOM organizes itself during an unsupervised learning process. For each training vector \mathbf{v} the winner weight vector \mathbf{w} is found. Then the winner and all his neighbouring nodes are adjusted with a learning rate α , which is decreasing from 1 to 0 during the training process:

$$w_{new} = w_{old} + \alpha(v - w_{old}) \quad (1)$$



(a) first stage: SOM trained on dataset containing mostly cosmic ray and satellite crossing types of bogus detection. (b) second stage : SOM trained on dataset of candidates appearing on at least 2 independent observations (c) third stage: SOM trained on dataset containing candidates which had been accepted by second-stage SOM (b).

Fig. 2: Visualizations of trained SOMs

3 Results

Using representative training datasets from OGLE-IV pipeline we trained three different SOMs to be used at subsequent stages of candidates verification, resulting altogether in a triple-stage verifier (Fig.2). Thumbnail images of candidates are being represented as vectors of pixel and their histogram values. In order to use SOM as a verification tool for transient candidates one has to choose which nodes form the acceptance region (i.e. images classified into them are verified as positives and accepted). The best to choose are nodes with highest percentage of true positives among all classified objects – we call it node’s purity, similarly to (Bloom et al., 2012).

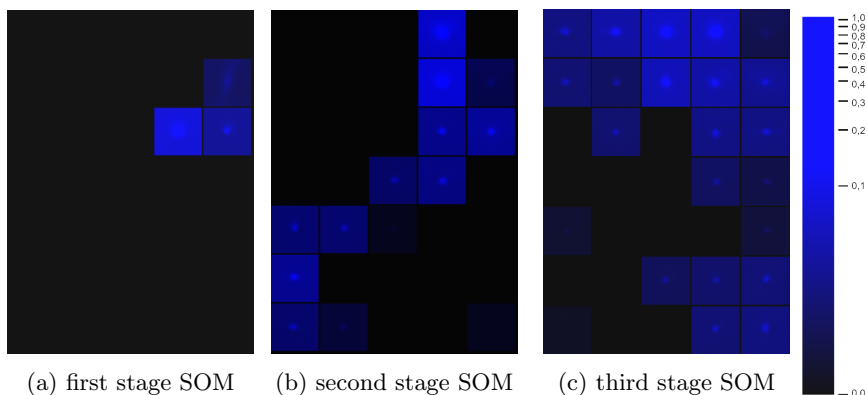


Fig. 3: Relative purity of nodes for each SOM, estimated based on test verifications of candidates of known types (real or bogus).

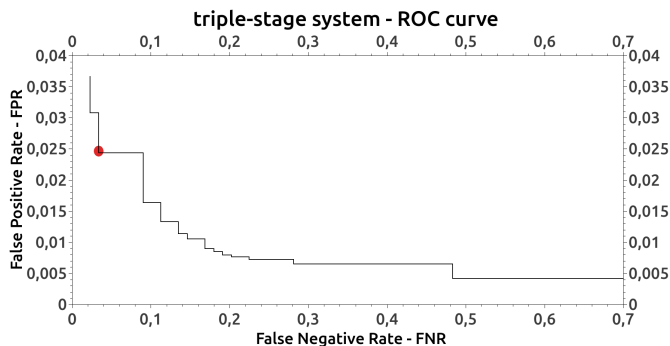


Fig. 4: Receiver operating characteristic (ROC) curve for the triple-stage system.

The receiver operating characteristic (ROC) curve for the triple-stage system (Fig.4) characterizes the verify performance depending on how many nodes are chose for the acceptance region. For an exemplary threshold marked with red dot the system accepts up to 97% of real transient detections and rejects 97.5% of bogus candidates.

References

- Bloom, J. S., Richards, J. W., Data Mining and Machine Learning in Time-Domain Discovery and Classification, 89–112 (2012)
- Bloom, J. S., et al., *Automating Discovery and Classification of Transients and Variable Stars in the Synoptic Survey Era*, PASP **124**, 1175 (2012), 1106.5491
- Kohonen, T., *Self-Organized Formation of Topologically Correct Feature Maps*, Biological Cybernetics **43**, 59 (1982)
- Udalski, A., Szymański, M. K., Szymański, G., *OGLE-IV: Fourth Phase of the Optical Gravitational Lensing Experiment*, Acta Astron. **65**, 1 (2015), 1504.05966
- Wyrzykowski, L., et al., *OGLE-IV Real-Time Transient Search*, Acta Astron. **64**, 197 (2014), 1409.1095